□□□□ HỌC THỐNG KÊ QUA TRUYỆN TRANH □□□□

CHƯƠNG 2: MÔ TẢ DỮ LIỆU

(tiếp theo)

Dữ liệu là nguyên liệu thô của thống kê, chúng tôi sử dụng các con số nhằm giải thích cho thực tế. Tất cả các vấn đề thống kê liên quan đến thu thập, mô tả, và phân tích dữ liệu, hoặc là suy nghĩ về việc thu thập, mô tả và phân tích dữ liệu.

Chương này tập trung vào mô tả dữ liệu. Làm thế nào để chúng ta có thể trình bày dữ liệu theo những cách hữu ích? Làm thế nào để chúng ta có thể nhận thấy mô hình cơ bản trong một đống các con số thông thường? Làm thế nào chúng ta có thể tóm tắt dạng cơ bản của dữ liệu?

Vâng, để mô tả dữ liệu, điều đầu tiên bạn cần là một vài dữ liệu thực tế... vì vậy chúng ta hãy cùng thu thập dữ liệu đã!

Dưới đây là một vài dữ liệu thực tế, đó là một phần của thử nghiệm trong lớp học. 92 sinh viên chính quy đã báo cáo trọng lượng của họ, kết quả như sau:

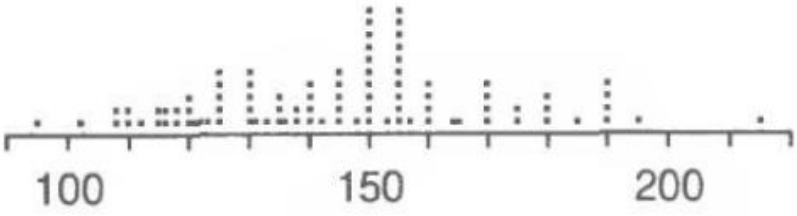
Nam

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180 135
170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160 135 160
130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

Nữ

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135 125
118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

Bắt tay vào việc cần làm, chúng ta vẽ được một đồ thị điểm: mỗi điểm trên đồ thị chính là cân nặng của một học sinh được báo cáo.



Cân nặng (pound)

Bạn có thể thấy vấn đề ở các nhóm có cân nặng 150 và 155 pound. Các học sinh có xu hướng báo cáo cân nặng của họ tăng thêm 5 pound. Trong tình huống thực tế, việc làm tròn số có thể che khuất mô hình dữ liệu chung... nhưng hiện tại, chúng ta mới chỉ khắc phục được hiện tượng này.

Chúng ta có thể tổng hợp dữ liệu với một bảng tần suất. Chia số liệu thành các khoảng và đếm số cân nặng của sinh viên trong mỗi khoảng. Tần suất là tỷ lệ của số cân nặng trong mỗi khoảng, bằng tần số chia cho tổng số học sinh.

Khoảng	Điểm giữa	Tần số	Tần suất
87.5-102.4	95	2	.022
102.5-117.5	110	9	.098
117.5-132.4	125	19	.206
132.5-147.4	140	17	.185
147.5-162.4	155	27	.293
162.5-177.4	170	8	.087
177.5-192.4	185	8	.087
192.5-207.5	200	1	.011
207.5-222.4	215	1	.011
Tổng		92	1.000

Lưu ý: Ta cần giữ các giới hạn khoảng cách xa những bội số 5 pound rắc rối. Điều này tránh được sai số hệ thống trong báo cáo của các học sinh…

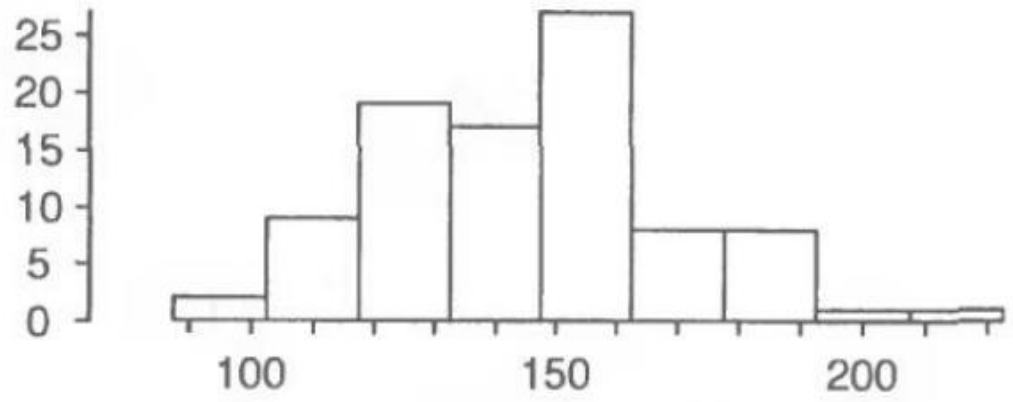
Quy trình tạo ra các khoảng cách tổ

1) Sử dụng các khoảng có độ dài bằng nhau với điểm chính giữa xung quanh các số thích hợp.

2) Đối với một tập dữ liệu nhỏ, sử dụng một số lượng nhỏ các khoảng.

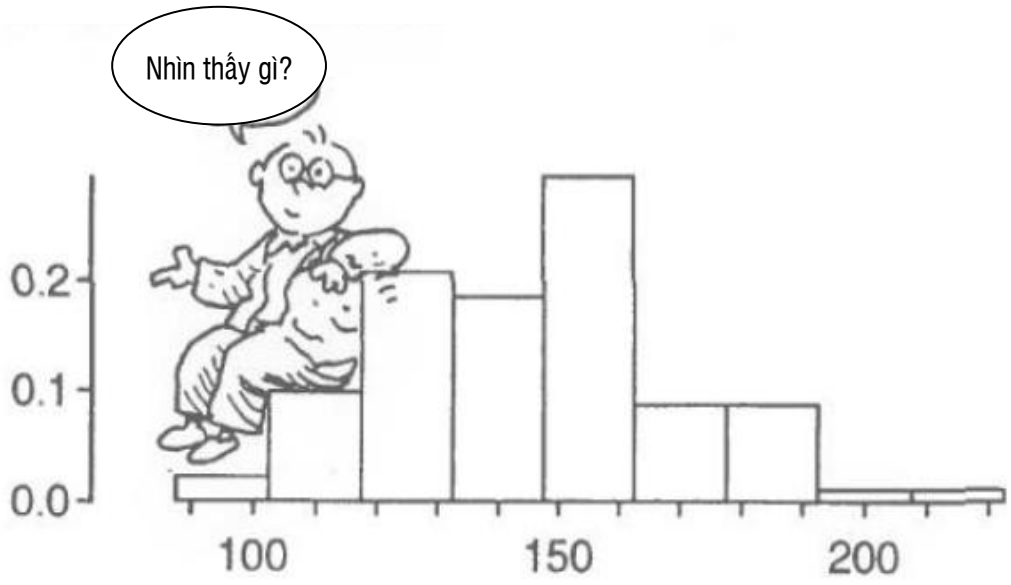
3) Đối với một tập dữ liệu lớn, sử dụng nhiều khoảng hơn.

Từ bảng tần số, chúng ta sẽ chỉ ra được có bao nhiêu dữ liệu điểm "bao quanh" mỗi giá trị. Chúng ta cũng có thể vẽ đồ thị cho các thông tin này. Đồ thị kết quả dạng cột được gọi là một biểu đồ. Mỗi cột bao gồm một khoảng và trung tâm là điểm chính giữa. Chiều cao của cột là số điểm dữ liệu trong khoảng.



Cân nặng (pound)

Chúng ta cũng có thể vẽ biểu đồ tần suất về cân nặng của học sinh. Hình dạng của biểu đồ tần suất giống hệt biểu đồ tần số, ngoại trừ quy mô theo chiều dọc.



Cân nặng (pound)

Nhà thống kê John Tukey đã sáng tạo ra một cách tổng hợp dữ liệu nhanh chóng vẫn giữ nguyên các điểm dữ liệu độc lập. Phương pháp đó được gọi là sơ đồ thân lá.

Đối với các dữ liệu cân nặng, thân là một cột các số, bao gồm các dữ liệu cân nặng tính bằng hàng chục (có nghĩa là chúng ta bỏ ra số cuối)

9
10
11
12
13
14
15
16
17
18
19
20
21

Bây giờ thêm chữ số cuối cùng của mỗi trọng lượng ở hàng thích hợp.

9 :
10 :
11 : 628
12 : 0155005
13 : 080015
14 : 05
15 : 0
16 :
17 :
18 :
19 :
20 :
21 :

Điền đầy đủ sẽ như sau:

9 : 5
10 : 288
11 : 628855060
12 : 01553005525
13 : 85008506000153
14 : 05505580502
15 : 505370550550505050050O
16 : 050004
17 : 055000
18 : 0500
19 : 00500
20 :
21 : 5

Và cuối cùng, đặt "lá" theo thứ tự

9 : 5
10 : 288
11 : 002556688
12 : 000123555555
13 : 0000013555688
14 : 000025555558
15 : 000000000003555555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

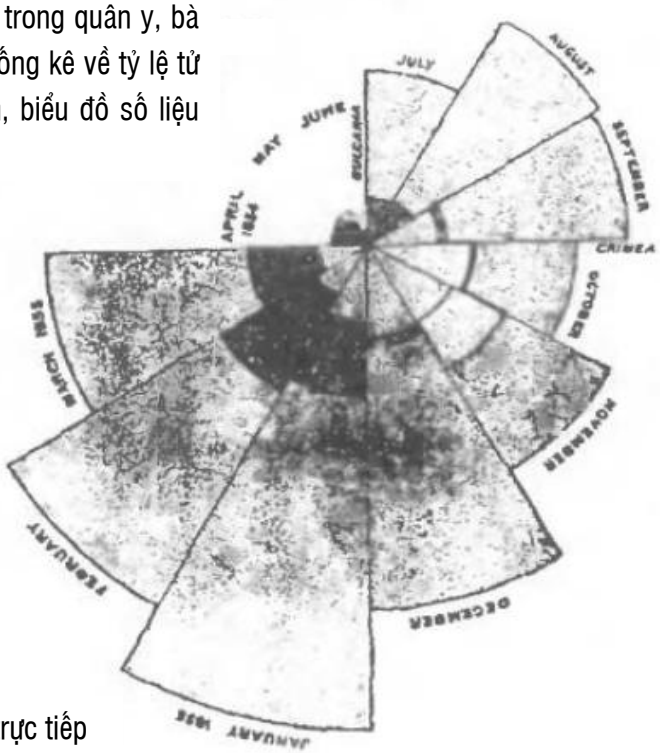
Tất cả những số 0 và số 5 rõ ràng cho thấy sự chệch trong báo cáo của học sinh!

Việc trình bày đồ thị tốt chính là một phần nghệ thuật và khoa học.

Đôi khi còn là một phần chính trị

Florence Nightingale là một y tá trong quân y, bà đã biên soạn được bộ số liệu thống kê về tỷ lệ tử vong ở bệnh viện quân đội Anh, biểu đồ số liệu đã gây bất ngờ như sau:

Các trục xuyên tâm chỉ ra các trường hợp tử vong (trong bệnh viện cũng như trên chiến trường) của binh sĩ Anh trong cuộc chiến tranh Crimean.



Những nỗ lực thống kê của bà đã trực tiếp cải thiện được tình hình của bệnh viện và giúp giảm tỷ lệ tử vong.

Tôi đã được cứu sống nhờ thống kê

(Còn nữa)

Biên dịch: Minh Ánh và các nghiên cứu viên, Viện Khoa học Thống kê